Probability and Exam Practice

Shannon Pileggi

STAT 217

OUTLINE

Probability

Exam Practice Questions

Probability

In 2013, Angelina Jolie got a double mastectomy because doctors told her that with the BRCA1 gene she has an 87% risk of breast cancer.

Would you get a double mastectomy under these circumstances? (If you are male, consider the question to be "Would you want your mother to get a double mastectomy under these circumstances?")

1. Yes
2. No

Probability

- Everyday you experience events in which the outcome is uncertain - these are **random phenomena**
  - There is a 70% chance of rain today
  - A new cancer treatment is successful for 40% of patients
  - The probability that I win the lottery is 1 in one million
- **Probability** is the way we *quantify* uncertainty or randomness.
- You have to *interpret* probability in **your everyday lives**.
- You have to *interpret* probability in **statistical analysis**.

Probability and Randomness

The **probability** of an outcome is the proportion of times that an outcome would occur in a long run of observations, or trials. Basic rules of probability are:

- A probability is always a number between 0 and 1.
- The sum of all of the probabilities for all the possible outcomes equal 1.

Sometimes probabilities are reported percents, in which case it should be between 0 and 100%.

Thought exercise

By yourself... here is a number line from zero to one that represents probabilities. Draw a cutoff point, and label it with a number, such that you classify probabilities as

- small - an event with this probability would be unusual to happen by random chance
- not so small - it would not be unusual for an event with this probability happen by random chance

Below is an example with 0.5 as a cutoff. Draw **your** cutoff point where **you** see fit.

Finding probabilities

- Make *assumptions* about your random process in order to calculate a probability (e.g., each roll of the die is equally likely)
- *Estimate* a probability with a sample proportion from a *long run* of observations (e.g., collect large amounts of data from which you an estimate a probability)
- *Estimate* a probability from simulating outcomes

The Monty Hall Problem

- In the game show Let's Make a Deal you choose one of three doors and win what is behind it.
- One door has a Cadillac and the two others have goats.
- The host knows where the Cadillac is and opens one of the doors you did not choose to reveal a goat.
- You are offered the chance to stay with your door or switch to last unopened door.

Do you have better chances of winning if you

1. stay
2. switch
3. either stay or switch has equal chance of winning

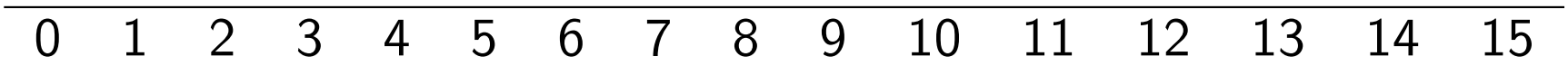
Play Let's Make A Deal!

- Pair up; assign one person to be the host and other to be the contestant. Get 3 index cards; write *goat* on 2, and *car* on 1.
- Simulation 1: Host shuffles cards, and can see prizes. Contestant selects a card; host reveals one card that is a goat; contestant employs *stay* strategy. Record prize won; repeat 15 times.
- Simulation 2: Host shuffles cards, and can see prizes. Contestant selects a card; host reveals one card that is a goat; contestant employs *switch* strategy. Record prize won; repeat 15 times.

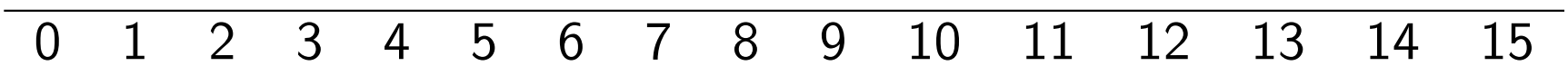
Repetition	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	# cars
STAY																
SWITCH																

Plot class results

Number of car wins under STAY strategy:



Number of car wins under SWITCH strategy:



Simulate Let's Make A Deal!

Open this website in internet explorer to do a *long run* simulation with *many* games.

http://www.grand-illusions.com/simulator/montysim.htm

Interpret the results

The probability of winning under the *stay* strategy is _____; the probability of winning under the *switch* strategy is _____.

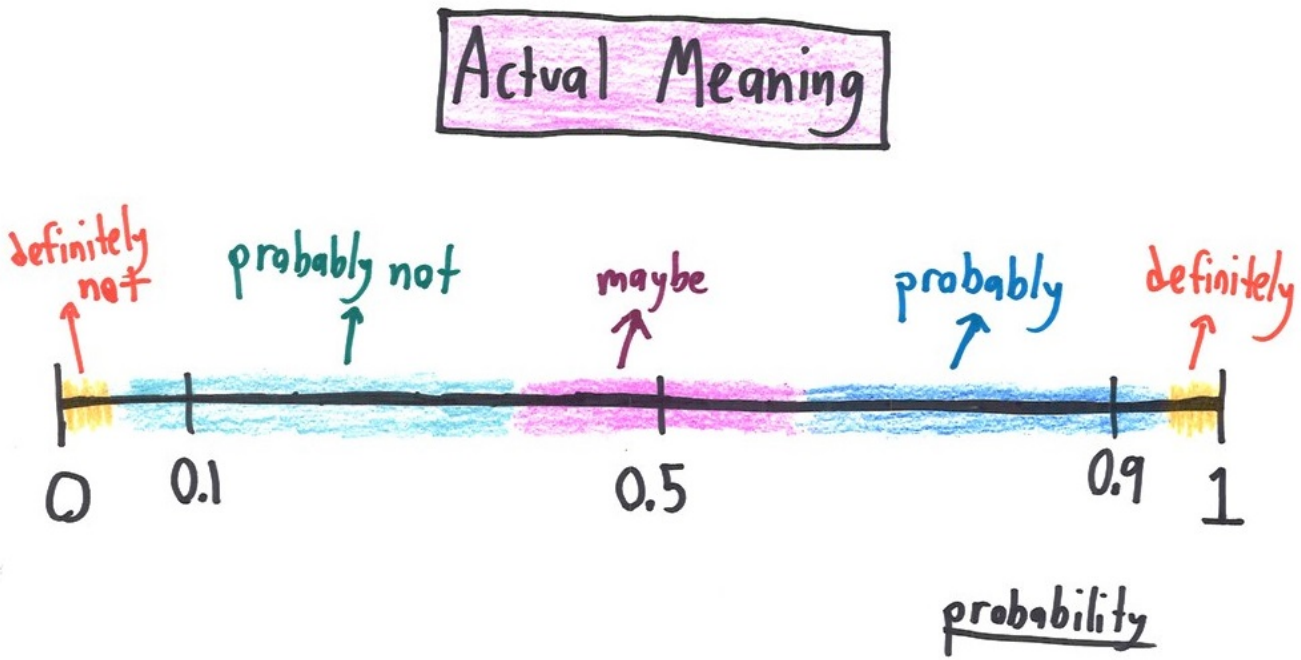
That is, if you play the game repeatedly under the same conditions, then after a very large number of games, your proportion of wins under the *stay* strategy should be very close to _____.

Group Exercise

Suppose a weather forecaster states that the probability of rain in SLO tomorrow is 0.30. Which of the following statements are *true*?

1. It will rain in 30% of SLO tomorrow.
2. It will rain 30% of the day tomorrow.
3. Out of 10 days with the exact same weather conditions as tomorrow, it would rain on exactly 3 of those days.
4. In the long run, among many days with the exact same weather conditions as tomorrow, it would rain on 30% of those days.
5. More than one statement is true.

Interpreting probabilities



Substitutions:

- "definitely not" = "highly unlikely"
- "definitely" = "highly likely"

Summary

Interpreting a probability is important in statistical inference. Remember the framework for statistical reasoning...

1. The data arose from random chance
2. The data didn't arise from random chance - something is really going on here

We assume the chance model is true, and then we determine how likely it is for our observed data to come from the chance model. When we determine that it is unlikely, or that the probability is 'small', we conclude that we have evidence that something is really going on.

Probability

Exam Practice Questions

A researcher asks 1000 families how many times a year they go out to eat.

Which sample statistic would be an appropriate measure of central tendency for the data?

1. the interquartile range
2. the standard deviation
3. the sample mean
4. the sample proportion

In a study about Facebook and researchers found that "using facebook made people feel worse about themselves." Study participants completed online surveys regarding their feelings as well as time spent on facebook over the course of the day.

Which of the following represents a variable in that study?

1. number of participants
2. they received 5 text messages a day
3. average life satisfaction score among all participants
4. how much facebook was used in a day
5. percent of females in the study
6. more than one is a variable

Data from a statistics class survey

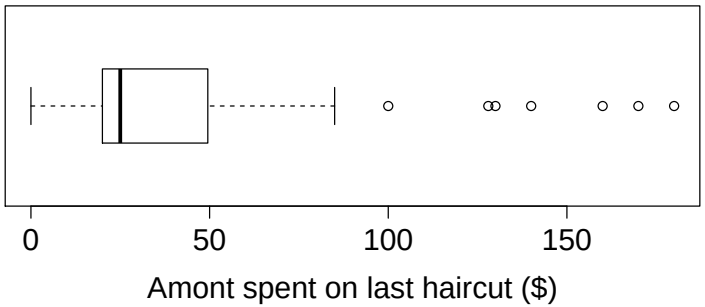
gender	sleep	bedtime	countries	dread
male	5.0	12-2	12	low
female	7.0	12-2	7	high
male	6.5	10-12	1	high
female	8.0	8-10	2	medium
female	7.3	10-12	9	high

gender	male/female
sleep	amount of sleep in a typical night (hours)
bedtime	time frame for nightly bedtime
countries	number of countries visited
dread	level of dread towards statistics (low, medium, high)

Which variables are quantitative?

1. sleep, bedtime, countries
2. sleep
3. bedtime, countries
4. dread, bedtime
5. sleep, countries

min	Q1	median	Q3	max	mean	sd	n
0	20	25	49.5	180	40.1194	41.59892	67



This data summary shows the distribution of amount spent on a haircut by 67 STAT 217 students. Which of the following statements is *true*?

1. This is a left skewed distribution.
2. Fewer students spent \$20-\$25 on a haircut than \$25-\$49.50.
3. 25% of students reported spending more than \$49.50.
4. 50% of students reported spending less than or equal to \$40.11.
5. More than one statement is true.

	More Strict	Less Strict	Total
Democrat	454	62	516
Republican	363	104	467
Total	817	166	983

Participants in the 2006 General Social Survey were asked if gun control should be stricter after the 9/11 tragedy and about their political affiliation. Which proportions should you compare if you want to determine if political affiliation is associated with views on gun control?

1. 817 / 983 vs 516 / 983
2. 516 / 983 vs 467 / 983
3. 454 / 516 vs 454 / 817
4. 454 / 516 vs 363 / 467
5. 454 / 817 vs 62 / 166

What is the *main* difference between observational studies and experiments?

1. Experiments take place in a lab while observational studies do not need to.
2. In an observational study we only look at what happened in the past.
3. Most experiments use random assignment while observational studies do not.
4. Observational studies are completely useless since no causal inference can be made based on their findings.

Students complain that a chemistry exam is too hard, while the professor says that the the exam is not too hard.

If exam scores are left-skewed, which measure are they using to describe 'typical' exam performance and justify their arguments?

1. the students are using the mean, whereas the professor is using the median
2. the students are using the median, whereas the professor is using the mean
3. both the professor and the students are using the mean
4. both the professor and the students are using the median

Historians use text analysis to attempt to attribute authorship of unknown works. From examining a body of known works of approximately 1000 words, author X uses 'thee' on average 14 times with a standard deviation of 3, and author Y uses 'thee' on average 20 times with a standard deviation of 2. The z-score for the unknown work relative to author X is 1.67, and the z-score for the unknown work relative to author Y is -0.5.

Which of the following statements is *true*?

1. the number of times 'thee' is used is more consistent with author X than author Y
2. the number of times 'thee' is used in the unknown work is 17
3. the unknown work uses 0.5 fewer "thee's" than typical for author Y
4. the unknown work uses 1 fewer thee than typical for author Y

A news story reported "Better fathers have smaller testicles" based on research by Emory anthropologist Dr. James Rilling. Biological fathers of children aged 1 or 2 years old who were currently cohabitating with the child's mother were recruited through using flyers posted around the Emory University campus, at local parks, daycare centers, and with an electronic advertisement on Facebook. Dr. Rilling used MRI scans to measure testes size and a self-report questionnaires to assess parenting involvement.

This is a _____ study, and therefore we _____ conclude that smaller testicle size causes men to be better fathers. _____ bias could have entered the study by the method of the participant recruitment.

1. observational, cannot, sampling
2. experimental, can, sampling
3. observational, cannot, response
4. observational, can, response

Scores on the verbal section of the SAT have a mean of 500 and a standard deviation of 100. Scores are normally distributed.

What proportion of verbal SAT scores are higher than 600?

1. 0.025
2. 0.05
3. 0.16
4. 0.32
5. 0.68

Suppose we are interested in the relationship between age and exercise habits. We randomly sample 3000 adults, and we collect information on their age and how many minutes a week they exercised.

Which figure would be most appropriate to begin to visualize if there is an association?

1. dot plot
2. histogram
3. scatterplot
4. side by side boxplot
5. barplot

Suppose that battery life of a laptop follows a normal distribution with a mean of 7 hours and a standard deviation of 2 hours. The 80^{th} percentile of of battery life is

1. less than 7 hours
2. greater than 7 hours
3. less than 2 hours
4. 7 hours
5. 2 hours
6. not enough information to determine